IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 9 February 2024, accepted 20 February 2024, date of publication 23 February 2024, date of current version 1 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369586

RESEARCH ARTICLE

Explainable Models for Predicting Academic Pathways for High School Students in Saudi Arabia

MAI ABDALKAREEM AND NASRO MIN-ALLAH
Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, Saudi Arabia

Corresponding author: Mai Abdalkareem (2210500248@iau.edu.sa)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Imam Abdulrahman bin Faisal University.

ABSTRACT The science of data mining has contributed considerably to the education sector. However, most educational data mining (EDM) studies have focused on predicting the future performance of students and detecting at-risk students to provide early targeted interventions. A few studies used machine learning techniques to predict the future academic pathways for degree students. However, there are limited studies that use the datasets of high school students to predict future pathways. Moreover, such studies are yet to be conducted using datasets produced in Saudi high schools. Therefore, researchers that work in the education sector have focused on EDM and are eager to apply advanced computer science methods to upgrade old administrative systems. Furthermore, the education sector is a rich field with data that can be exploited to improve and enhance educational management systems and accelerate digital transformation. In this study, we explore the applications of EDM and review the algorithms used by other researchers in this field. We applied supervised machine learning classifiers to educational datasets collected from high schools in Saudi to predict the future academic pathways of students and identify the essential factors that affect them. This study contributes to the literature by developing a predictive model for students in Saudi high schools and detecting critical features that affect future academic careers of the students. Furthermore, we used explainable artificial intelligence to interpret the best model and enhance its transparency.

INDEX TERMS Classification, data mining, educational data mining, explainable artificial intelligence, feature selection, machine learning, Shapley additive explanations (SHAP) value, student.

I. INTRODUCTION

During the COVID-19 pandemic in Saudi Arabia (SA) in 2020, more than five million students, approximately 450 thousand faculty members, and more than one million parents or guardians used the Madrasty platform. The platform registered 489 million online visits by the end of the 17th week of the first semester. Further, teachers had performed up to 89 million synchronized virtual classes and generated more than 15 million homeworks for their students in that short period. Statistics released by the Saudi Ministry of Education's Twitter account indicate that these numbers have considerably increased [1].

From a computer science perspective, a tremendous amount of generated and stored data can be exploited to enhance and improve the educational processes in SA. Thus, in this era of big data, more advanced methods (such as data mining (DM) techniques) are recommended to exploit the enormous amount of educational data to benefit students, teachers, and the educational environment.

Recently, data have been considered an essential asset in making informed decisions. Therefore, the application of advanced methods is beneficial for extracting targeted useful knowledge. Machine learning (ML) is a part of artificial

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

intelligence that aggregates all methods that allow a machine to learn and perform accurate predictions based on past observations [2].

Educational data mining (EDM) is a new discipline that applies DM techniques (i.e., ML and deep learning (DL) algorithms) to educational data to extract valuable knowledge, detect patterns, and identify effective related features to better understand student behavior. However, EDM transforms the raw data extracted from learning management systems into beneficial information that can enhance educational research and systems [3], [4]. EDM is the intersection of three main fields: education, statistics, and computer science [2], [3], [4], [5]. Three subsections were generated at the intersection of three fields: DM and ML, learning analysis, and computer-based education. The EDM process is illustrated in Fig. 1.

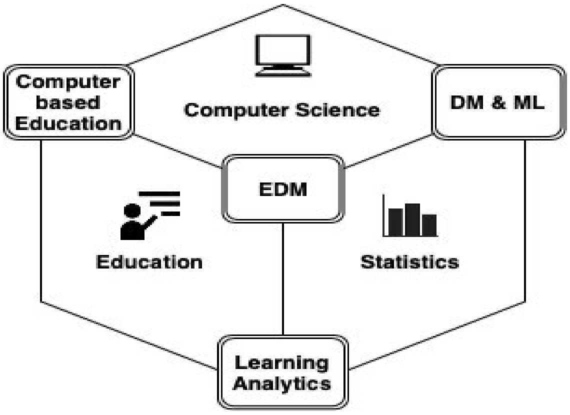


FIGURE 1. Educational data mining.

Educational datasets have been conventionally analyzed using DM techniques (such as classification, clustering, text mining, and association rule mining) [4], [5], [6], [7]. Classification and regression are supervised learning techniques. These techniques aim to build models based on the observed data and well-known association results [2]. Therefore, in these methods, the datasets must contain a labeled class to train the algorithms. Thus, the data were divided into training and testing datasets. The algorithms were then trained using the training set and the test set was applied.

The primary difference between classification and regression was the class label category. The class label is categorical in classification; additionally, it can be binary, multiclass, or multilabeled. On the contrary, it is numerical in regression, and the models predict missing continuous values.

Clustering is an essential unsupervised DM method. This type uses only input observation data, without association outputs. Hence, the data were grouped into similar clusters by determining hidden patterns and similarities in their characteristics.

This study has two motivations for focusing on EDM. First, researchers in the education sector are eager to apply advanced computer science methods to upgrade the old administrative systems. Second, the education sector is a rich field with data that can be exploited to improve and enhance educational management systems and to accelerate digital transformation.

Schools and universities have generated and saved enormous amounts of student data using education learning management systems. These enormous amounts of data can be transformed into valuable knowledge to guide decision makers in the Ministry of Education and even lead students to succeed in their future academic pathways. Thus, EDM techniques such as ML and DL algorithms can be applied to educational datasets to extract valuable knowledge, detect patterns, and identify the most relevant features.

Applying DM techniques to educational contexts can help students and faculty members in the education sector by predicting students at risk of failure and identifying the crucial factors affecting their success. Administration in this sector can utilize this model in the early preparation and scheduling of the next academic year. Students will also become aware of the factors related to their academic success.

In this study, we applied supervised ML classifiers to educational datasets collected from Saudi high schools to predict the future tracks of students and identify essential factors affecting their future pathways.

The main objectives of this study are as follows:

1. Identify and collect features significant to the future pathways of students.

2. Develop multiple ML predictive models and select an effective model for predicting academic pathways among Saudi high school students.

3. Analyze and identify the most related factors influencing the academic pathways of students.

The research questions of this study are as follows:

RQ1: Is it possible to predict the future academic path of Saudi Arabian high school students?

RQ2: What are the most related features influencing the academic pathways of students?

RQ3: How can the interpretability of EDM models be utilized to increase their transparency?

The remainder of this paper is organized as follows. Section II presents a review of the literature related to EDM, and Section III presents the proposed methodology. Section IV presents the discussion and findings. Lastly, Section V presents conclusions and future research directions.

II. RELATED WORK

Here, we present the most recent studies related to EDM that apply DM techniques to an educational context. Various researchers have explored and reviewed recent studies (published in the last six years) related to EDM and its applications to better understand the behaviors of students and education systems. Subsequently, we addressed the DM, ML, and programming tools used in these studies.

After reviewing 27 related studies, four main objectives were identified. First, most studies used classification and regression approaches to predict student performance. One study was conducted to predict the attention of students toward postgraduate programs. Only a few studies have used feature selection methods to rank the top features affecting student performance. Finally, only a few studies on predicting students' academic pathways have used the clustering method.

A. DIFFERENT STUDY OBJECTIVES OF EDM

1. Predicting the performance of students.

2. Predicting the interest of students in postgraduate programs.

3. Feature selection to identify the most relevant features.

4. Predicting the academic pathways of students (academic programs).

A comprehensive summary of each objective and related study is presented below: These studies are arranged and classified based on the similarity in approach to this study in the first subsection, and the similarity in objectives in the second and third subsections.

1) PREDICTING THE PERFORMANCE OF STUDENTS

Here, we review studies that have used classification and regression methods to predict students' performance.

Kuzilek et al. conducted a qualitative pedagogical study [8]. They observed a drawback in the performance of students in their first academic year at a Czech University. Thus, they utilized their ML knowledge to build a predictive model that could predict the success of students in their first academic year. They conducted three experiments, and the results indicated that sequential data-based models with a length of two sequences achieved the best performance in detecting failing students. The enhancement was significant compared to baseline predictions, which only used examination results, given that the cumulative state data were richer than the examination results data. However, the detection of passing students was more accurate in the sequential and baseline predictive models. Moreover, basic behavioral patterns were detected in the data. These results can help faculty intervene and provide targeted help to at-risk students.

Hashim et al. proposed another EDM model to discover hidden functional patterns and explore meaningful information from educational data [9]. The training dataset for the supervised ML techniques included the main factors, such as demographics, academic grades, and behavioral features. Courses in bachelor's programs at Basra University provided the dataset used to train this model. Various supervised ML methods (i.e., naïve Bayes (NB), logistic regression (LR), K-nearest neighbor (KNN), and artificial neural network (ANN)) were implemented in this experiment. The experimental results indicated that the LR classifier outperformed the other classifiers in predicting the final grades of students, with 68.7% and 88.8% accuracies for passed and failed, respectively.

To improve the educational quality of students, Karthikeyan et al. [10] proposed a new and efficient hybrid EDM model (HEMD) to analyze the performance of students. Distinctive factors were used to evaluate students' performance in delivering precise results. Thus, NB and J48 classifiers were combined to specifically categorize student performance. Additionally, Weka environments were used for the evaluation process with a dataset available online. A comparative study compared HEMD with other existing approaches and proved that the accuracy rate of the proposed model was enhanced by 10%, reaching 98%.

To maximize the contribution in EDM, Sokkhey et al. used multiple models to maximize the contribution of EDM [11]. Data were collected from various high schools, and ML, DL, and statistical analysis techniques were applied to predict mathematical performance. A spot-checking algorithm is used to determine the most effective technique. The results demonstrate that the random forest (RF) classifier achieved the highest accuracy of approximately 97%. Moreover, they introduced a feature selection method to determine the most relevant features that affect student performance. They identified the top ten essential features, with the top three related features being students' interest in the course, spending hours on self-study, and doing homework frequently.

Saleem et al. [12] used a dataset extracted from an electronic learning management system with several features to predict student performance. They proposed an integrated ML model to help make intelligent and proactive decisions based on the evaluated students' performance. They used decision trees (DT), RF, gradient boosting trees (GBT), NB, and KNN in their experiments; however, they generated insufficient results. Hence, applying bagging, stacking, voting, and boosting ensemble techniques enhances the performance. The stacking model that aggregated the five classifiers outperformed the other ensemble methods and achieved the highest F1-score (82%).

Harvey and Kumar [13] worked on math score prediction scholastic aptitude test for high school students. The proposed model included three classifiers: linear regression, DT, and NB. NB outperformed the other classifiers, showing a 71% accuracy rate. The dataset was extracted from the Massachusetts State website and has many features (i.e., numeric grades, teacher salaries, school finances, and demographics).

Amra and Maghari [14] applied NB and KNN classifiers to predict the grades of high school students using a Ghaza strip. Although they used fewer algorithms than other studies, they focused on comparing these algorithms. The experimental results demonstrated that NB achieved a higher accuracy rate than KNN. NB had an accuracy rate of 93.17%.

Priya et al. [15] suggested a comprehensive EDM scheme that can predict student achievement and illustrated its potential factors. The proposed model examines psychological factors, study characteristics, and demographic data to aggregate information on students, teachers, and parents. LR, support vector machines (SVMs), and neural network classifiers were used in this study. LR outperformed the other classifiers and achieved an accuracy rate of approximately 68%. They claimed that their proposed framework outperformed other existing frameworks, whereas the three classifiers had similar accuracy rates.

Sunday et al. [16] collected student log data from a computer science unit at a university in Nigeria to analyze the performance of students in programming courses. An experiment was conducted to compare the J48 DT and ID3 classifiers in this study. The results demonstrate that the J48 algorithm had a higher accuracy rate (87.02%) than other classifiers. Moreover, information gain and gain ratio were used to select the essential features that affect student performance. They concluded that class attendance was the feature that was most closely related to student success.

Ahmed and Mahmood [17] implemented the decision table, RF, random tree, OneR, and J48 classifiers in Weka's open-source environment to predict and classify students' future grades. The experimental results demonstrated that J48 had the highest prediction accuracy compared with the other models, with a 78% accuracy rate.

Buenaño-Fernández et al. [18] proposed the application of ML algorithms that used students' historical grades to predict their final grades. The dataset was extracted from a university in Ecuador with a certain degree. The study was conducted in two phases. Students with similar performance patterns were divided into two groups. Next, an appropriate supervised ML technique was selected, and the experiment was conducted in a Weka environment. Finally, the researchers confirmed the effectiveness of the ML methods in predicting student performance.

Tarik et al. [19] implemented a referral system that enhanced guidance schemes by forecasting student performance in high schools. Three ML regression algorithms are used in this scheme: LR, DT, and RF. The dataset was extracted from an educational institute in Morocco and contained students' grades during the first and second years of high school. However, DT and LR were weak and insufficient for predicting student grades. The RF algorithm achieved the highest accuracy and predicted the best average score.

Kenekayoro [20] used an SVM to forecast the performance of students in two academic disciplines based on their historical grades in three subjects. Four feature selection techniques were used to study the relationship between the past knowledge of certain subjects and a particular discipline. The experimental results demonstrated that the SVM achieved up to 80% accuracy, illustrating that this technique can be developed to build a real-time recommender system. Moreover, the feature selection results demonstrated the following findings: First, this approach demonstrated no apparent correlation between scores of an individual subject and future performance in another discipline. Second, students with good average scores on previous subjects may perform well in the future studies. Ultimately, the results indicate that a strong background in mathematics is necessary to achieve above-average grades on the mathematics path.

Most ML-related studies have focused on datasets with extensive records and broad attributes of each student. Wakelam et al. [21] conducted an experiment involving 23 students with minimal attributes. They applied KNN, regression DT, and RF to predict the assessment grades of the students using only 10 attributes, with RF and KNN showing promising results with accuracies of 70% and 74%, respectively. Furthermore, decreasing the number of attributes produced mixed results, contrary to the prediction accuracy when all available features were used. Hence, faculty members cannot reliably account for student interventions. They found the potential to predict individual student's midterm and final exam grades in small student cohorts with minimal features.

Qazdar et al. [2] used multivariate regression ML techniques to develop two models to predict the students' results in the first semester and national exam. They claimed that the two models could predict the performance of students more precisely, although the accuracy of these models was not specified.

Alboaneen et al. [22] developed a web-based system to predict student performance using five regression-ML techniques. SVM, RF, LR, ANN, and KNN were applied to a dataset comprising of 10 features for 168 students. The dataset was collected from the computer department of Imam Abdulrahman University, and it included demographic and academic features. Two evaluation metrics—mean absolute error and percentage error—were used. LR outperformed all the techniques used, scoring 6.34%, which was the lowest mean absolute percentage error.

Sahlaoui et al. [23] proposed an ML approach to enhance the performance of a previous study using the Jordan dataset [24]. The default design of the synthetic minority oversampling technique (SMOTE) was used along with ensemble methods. Moreover, K-fold cross-validation was used to divide the dataset into training and test sets, with an optimal K value of 10. Furthermore, a hyper optimization process using a simple grid-search technique was used for parameter tuning. These processes resulted in an enhancement of over 20% in comparison to previous studies, and the model achieved an accuracy rate of more than 99%.

The uncertainty level in the ML model results decreased in the case of an unbalanced dataset. Bujang et al. [25] developed a predictive ML model that could increase the confidence level of the results of an imbalanced dataset. They proposed a multiclass predictive model using six ML algorithms: DT (J84), SVM, NB, KNN, LR, and RF. Notably, the performance of most classification models was improved when oversampling was performed using SMOTE with wrapped- and filter-based feature selection methods. However, the obtained results demonstrated that RF had the highest accuracy and an F-measure of 99.5% for both evaluation metrics.

Mengash [26] used four classification techniques with three attributes to determine students' performance before admission. The three attributes included three pre-admission criteria: general aptitude test score, average high school grade, and scholastic achievement admission test score (SAAT). ANN outperformed DT, SVM, and NB and had the highest accuracy rate (∼80%). The results also revealed that the SAAT pre-admission criterion was the most accurate for predicting student performance. Therefore, assigning more weights for SAAT in the admission system is recommended.

Nabil et al. [27] used EDM to explore the effectiveness of DL in this field. A university dataset was used to build predictive models of student performance. DT, RF, gradient boosting (GB), LR, SVM, and KNN were utilized in this study and compared with deep neural networks (DNNs). Various resampling methods have been used to solve the imbalanced dataset problems. However, the experimental results demonstrated that the DNN achieved an 89% accuracy rate for both the imbalanced dataset and the SMOTE oversampling techniques. The RF outperformed the other models in the imbalanced dataset and achieved an asymptotic accuracy rate of 88% in the second case.

Adnan et al. [28] constructed predictive models using one DL and six ML algorithms to predict students at risk for course lengths of 0, 20, 40, 60, 80, and 100%. The main objective of this study was to characterize the behaviors of students based on their study features. The RF outperformed the other algorithms, with an average precision of 92% at a course length of 100%. Note that the (Course length) is a feature that has possible 6 values: 0,20,40,60,80 and 100. Hence, 0 for the students registered for the course but never attend any class. 20 for the students registered for the course but only attend 20% of the course length. 40 for the students registered for the course but only attend 40% of the course length. 60 for the students registered for the course but only attend 60% of the course length. 80 for the students registered for the course but only attend 80% of the course length. 100 for the students who attend the whole course.

Al-Shabandar et al. [29] proposed two models to detect learners at risk of dropping or failing courses and used a statistical method and four ML algorithms to train these models. The experimental results demonstrated that all classifiers achieved good accuracy rates in both models, whereas GB achieved the highest accuracy of 95.2% for the learning achievement model. Student motivation trajectories were the main reason for students withdrawing from e-learning.

2) PREDICTING THE ATTENTION OF STUDENTS TOWARD POSTGRADUATE PROGRAMS

Here, we review the prediction of students' attention toward postgraduate programs, even before applying for a degree.

Lin et al. [30] predicted students' intentions to attend master's programs after graduation or to find a job. They developed a new model that aggregates a modified fuzzy KNN, RF, and a novel chaos-enhanced sine-cosine algorithm (CESCA). The RF was used in this model to evaluate the importance of the features. The key parameters of the fuzzy KNN are automatically tuned using CESCA. The experimental results demonstrated that the proposed model could predict the intentions of students even before applying for a master's degree, achieving an accuracy of up to 82%.

3) FEATURE SELECTION TO IDENTIFY THE MOST RELEVANT FEATURES

Here, we review the studies that have used the feature selection method. Although these studies were conducted to predict the performance of students, they used a feature selection method to rank the most related features that affected the performance of learners.

Liu et al. [31] proposed a framework that uses student behavior and exercise features and aggregates the attention mechanism with a knowledge tracing model to predict the performance of students. First, they exploited ML to automatically capture feature representations. Second, they used fusion attention techniques based on a recurrent neural network architecture to predict student performance. The efficiency and effectiveness of this approach were proven by experimental results on a genuine dataset. Their model outperformed the previous models, with a high accuracy rate of 98%.

Polyzou and Karypis [32] aimed to identify at-risk students performing worse than usual, failing, or dropping a course before taking a specific course. Important features that can identify at-risk students were successfully extracted from the historical grading data by applying simple and sophisticated ML classifiers. Among the RF, DT, SVM, and GB classifiers, GB exhibited the best performance based on two metrics: the area under the curve (AUC) and maximum F1-score. They concluded that the performance of a single model can be enhanced by combining the predictions of different models.

4) PREDICTING THE ACADEMIC PATHWAYS OF STUDENTS

Here, we review two studies that used the clustering approach to predict students' academic pathways.

Regarding the identification of learning pathways for university students, Iaterllis et al. proposed a new ML model [33]. An unsupervised clustering algorithm was applied to detect students at risk of academic failure and to predict their future careers. The k-means clustering algorithm was applied to a dataset covering three academic years from a computer science program at a Greek university. This study divided students into three distinct areas of specialization based on outcomes, similar characteristics, and education-related factors. Quality standards, research work, time required to complete the degree, and degree completion time were selected for the clustering model. However, by choosing parameter K to be greater than three, the Euclidean distance formula produced significant clusters.

Iaterllis et al. [34] predicted student paths using a new technique. The new method has two phases and takes advantage of both the supervised and unsupervised ML algorithms. The experimental results of a previous case study using k-means clustering revealed that the dataset had three

TABLE 1. Literature review classifies according to domain and scope.

Reference	Country (Domain)	Scope
[2]	Morocco	High school
[8]	Czech	University
[9]	Iraq	University
[10]	Portugal	High school
[11]	Cambodia	High school
[12]	-	E-learning
[13]	USA	Elementary and high school
[14]	Palestine	High school
[15]	Portugal	High school
[16]	Nigeria	University
[17]	Pakistan	University
[18]	Ecuador	University
[19]	Morocco	High school
[20]	Nigeria	University
[21]	-	University
[22]	Saudi Arabia	University
[23]	Jordan	Schools
[25]	Malaysia	Polytechnic
[26]	Saudi Arabia	University
[27]	Egypt	University
[28]	UK	University (E-learning)
[29]	UK	University (E-learning)
[30]	China	University
[31]	-	-
[32]	USA	University
[33, 34]	Greece	University

coherent clusters. Second, to address each cluster individually, the discovered clusters were used to train the prediction models. Subsequently, two ML models were trained for each cluster to predict the time required to complete student enrollment in a degree program. Three criteria were used to evaluate the model: precision, recall, and accuracy. The experimental results demonstrate that the accuracy rate of the clustering-guided method increased from 77% for the non-clustering-guided model to approximately 80%, whereas recall increased from 72% to 84%.

From a review of recent studies, it is evident that most previous EDM studies were conducted to predict the future performance of students and identify at-risk students to provide early targeted interventions. However, few studies have used ML techniques to predict future academic disciplines of degree students. A study was conducted to predict students' attention toward pursuing postgraduate studies or finding jobs.

Table 1 shows that the reviewed studies are classified according to domain and scope to highlight the spatial knowledge gap. It demonstrates that there are just two studies that used data from university students in Saudi Arabia. Some studies also used data from high school pupils, however their domains were distinct. For this reason, this study is the first that predicts Saudi high school students' academic paths.

The literature reviews are summarized in Table 2.

B. INTERPRETATION OF ML MODELS

Here, the outcomes of the ML model were interpreted and explained to provide a more comprehensive understanding of the results. In some cases, the interpretability of ML models is more crucial than their accuracy [23], [35]. A systematic review [36] highlighted the importance of explainable models in educational contexts and the gaps in the field of explainable models in educational studies. We conclude that the explainability of ML models has yet to be explored in educational contexts.

The Shapley additive explanation (SHAP) value can be used to enhance the transparency of models, as it helps determine the main reason for the decrease in the accuracy of the model. This powerful visualization tool can provide deep insight into the attributes that affect a model, which can be immediately interpreted by nontechnical users [37].

Sahlaoui et al. [23] used the SHAP value to explain the model, which resulted in a 20% enhancement in the achieved accuracy. Additionally, they revealed that students' absences had a significant effect on predicting their performance compared to other factors. Therefore, the use of SHAP values in ML models is recommended to enhance their performance and transparency.

The science of DM has proven its significant contribution to the education sector in predicting the scores of students, identifying the most influential features of students' success, predicting the attention of students toward postgraduate

TABLE 2. **Summary of literature reviews.**

Ref.	Year	Attributes	Dataset	Preprocessing	Algorithms	DM Techniques	Tools	Main objective	Results
[2]	2019	15	Collected from the Scholar Management System for high schools in Morocco	-	Multivariate regression and ML algorithms	Regression	Python	Predicting student performance	The accuracy rate of the two models is not mentioned precisely
[8]	2021	25 58 80 85 65 26	Two datasets were extracted from two consecutive academic years- six sequential datasets	-	KNN Classification and regression tree (CART) RF Nonlinear SVM with a radial basis function kernel (SVM–RBF)	Classification - Analyzing frequent patterns	-	Predicting a student's success	-Detection of passing students was more accurate in both sequential and baseline predictor - Basic exam behavior patterns were detected within the data
[9]	2020	Demographic, academic background, and behavioral features	Two academic years	Normalization	Supervised ML DT -NB-LR-SVM KNN Sequential minimal optimization Neural network	Classification	WEKA	Predict student performance	LR classifier outperformed other classifiers in forecasting the exact final grades of students, with 68.7% and 88.8% for passed and failed, respectively.
[10]	2020	14	The data set was obtained from Cortez and Silva [35] Collected from high schools in Portugal	Selection Transformation Normalization	HEMD NB Classification J48 classifier	Classification Multiclass	WEKA	Predict student performance	Enhanced accuracy rate by 10% in comparison with the previous approaches
[11]	2020	43	Obtained from many high schools	Normalization Transformation Discretization Dimensional reduction	Statistical analysis (ANOVA) ML algorithms Deep learning	Classification Multiclass	R	Predict student performance in math. Identify the top-affected features	RF achieved the highest accuracy, around 97% Students' interest in math is the most affected feature
[12]	2021	17	Kalboard 360 dataset - Kaggle	Selection	DT, RF, GBT, NB and KNN	Classification Multiclass	Raped Miner (RM)	Predict student performance	The stacking model outperformed and scored the highest F1 score, 82%

TABLE 2. ***(Continued.)*** **Summary of literature reviews.**

[13]	2019	27	Extracted from the Massachusetts state website	Normalization	NB-DT – Linear Regression	Classification	R	Predict student performance	NB outperformed – 71%
[14]	2017	8	High school grades of Secondary General Certificate	Selection	KNN-NB	Classification Binary class	RM	Predict student performance	NB outperformed – by 93.17%
[15]	2021	33	Collected from two secondary schools in Portugal	Normalization Selection	LR- SVM- Neural Network	Classification	-	Predict student performance	LR outperformed – around 68%
[16]	2020	6	Collected from one university in Nigeria	Selection Transformation	J48 DT- ID3	Classification	WEKA	Predict student performance	J48 DT had a higher accuracy- 87.02%
[17]	2018	9	Extracted from the Iqbal Open University database	Normalization	DecisionTable-OneR—J48-RF- Random Tree	Classification	WEKA	Predict students' performance	J48 had a higher accuracy rate – 78%
[18]	2019	8	Extracted from the academic management system of a university in Ecuador	Normalization	DT	Classification Identify patterns.	WEKA	Predict students' performance	DT: accuracy rate 96.5%
[19]	2021	11	Data extracted from the educational academy in Morocco	Normalization	LR- DT- RF	Regression	Python	Predict students' performance	RF had the highest average accuracy rate
[20]	2018	53	Collected from one university in Nigeria	Feature selection- Dimension reduction	SVM	Classification Binary class	Python	Predict students' performance	SVM: up to 80% accuracy rate
[21]	2020	10	University final-year students, a cohort of 23	-	DT- KNN-RF	Regression	-	Predict students' performance.	The average prediction accuracy achieved by RF was 75%
[22]	2022	Ten academic and demographic features	Collected from the computer science department at IAU and through a questionnaire	Normalization	SVM-RF-LR-ANN-KNN	Regression	Python Excel	Predict students' score	LR scored the lowest MAPE: 6.34%

TABLE 2. *(Continued.)* **Summary of literature reviews.**

[23]	2021	17	Jordan dataset-Collected from a Jordanian school	SMOTE	K-Neighbor Classifier (KNC) Bagging Classifier (BC) DT-RF Extra tree (ET) X Gradient Boost (XGB)	Classification Binary class	Python	Predict students' performance	BC and XGB outperformed other classifiers. Accuracy = 99.4%
[25]	2021	10	Collected by one of Malaysia's polytechnics	SMOTE Wrapper Filter	DT(J84) SVM NB KNN LR RF	Classification Multiclass	WEKA	Predict students' final grade	RF outperforms, F-measure = 99.5% and Accuracy = 99.5%
[26]	2020	3	Collected from a Saudi university	Normalization	DT SVM NB ANN	Classification Multiclass	WEKA	Predict students' performance	ANN has the highest accuracy rate, more than 79%
[27]	2021	12	Collected from a university	Discretization SMOTE ADASYN ROS SMOTE-ENN	Deep Neural Network DNN DT RF GB LR SVM KNN	Classification	Anaconda software (Spyder)-Python	Predict students' performance.	DNN outperformed with an accuracy of 89%
[28]	2021	13	Open University Learning Analytics Dataset (OULAD)	Normalization	RF SVM KNN Extra Tree AdaBoost GB	Classification Multiclass	Python	Predict students' performance.	RF outperformed with an average precision of 92% at 100% of course length
[29]	2019	13 14	OULAD dataset Harvard University dataset	Normalization Transformation SMOTE Feature selection	RF Neural networks GB Statical method (generalized linear model)	Classification	-	Predict students' performance.	GB outperformed for all features included with 93.3% AUC
[30]	2019	12	Collected from a Chinese university	Normalization	RF Fuzzy KNN (FKNN) CESCA–FKNN	Classification Binary class	MATLAB	Predict students' intentions for master's programs.	CESCA-FKNN scored an accuracy rate of up to 82%
[31]	2020	24	Online dataset extracted from the free learning platform	Normalization	-ML algorithms -DL: fusion attention mechanism based on neural network Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing (MFA-DKT)	Classification	Pytorch	Predict student performance	Up to 98% accuracy rate

TABLE 2. ***(Continued.)*** **Summary of literature reviews.**

[32]	2019	42	Students' grades from the University of Minnesota	Normalization wrapper	RF- DT- SVM- GB	Classification Binary class	Python	Predict student performance	GB had the best-performing results based on two metrics: Area under curve (AUC) and max F1-score
[33]	2020	Not explicitly discussed in this article - 13, as mentioned in the following article	It covers three academic years for students in the computer science program in Greece	Normalization	K-means clustering	Clustering	Python	Predict students' academic program	k>3 produced three coherent and well-divided clusters
[34]	2021	13	It covers three academic years for students in the computer science program in Greece	Normalization	K-Means clustering RF	Clustering Regression Binary classification	Python	Predict students' academic program	The accuracy rate for cluster guided approach was approximately 80%

Note: Acronyms used in Table 1: DM: Data Mining, ML: Machine Learning, DL: Deep Learning, KNN: K-Nearest Neighbor, SVM: Support Vector Machine, DT: Decision Tree, NB: Naïve Bayes, LR: Logistic Regression, GBT: gradient boosting tree, SMOTE: synthetic minority oversampling technique, BC: bagging classifier, ET: extra tree, XGB: extreme gradient boost, XAI: explainable artificial intelligence, ANN: artificial neural networks, TP: true positive, TN: true negative, FP: false positive, FN: false negative

studies, and predicting the academic programs of students. Here, we explore the applications of DM and review the various algorithms used by researchers in other educational sectors.

Most EDM studies in this study were conducted to predict student performance. We also established that supervised ML such as classification and regression is a common type of DM used in EDM studies. Python and WEKA are commonly used. RF, BC, X-gradient boost (XGB), DT, and NB are ML algorithms that achieve the highest accuracy rates. Notably, the interpretability of ML is rarely used in EDM although it contributes to the transparency and accuracy enhancement of models.

Here, we clarify other areas of EDM that require further investigation. In the following section, we use ML algorithms to predict Saudi high school students' academic pathways. Moreover, future EDM studies should consider the interpretability of ML models because it is crucial to enhance their transparency.

III. METHODOLOGY

Herein, we describe the methodology used in this study. We named it Masarat predictive models (MPMs) study. (Masarat is the Arabic word that means academic pathways.)

To achieve the research objectives, we answered the research questions in seven steps: defining the problem, data collection, data preprocessing and description, data modeling, optimization, verification and evaluation, and explaining and interpreting the models. These steps are comprehensively and chronically addressed and listed.

A. DEFINING THE PROBLEM

Until 2021, students in high schools in Saudi Arabia who accomplished their first mutual year could study science or art paths for the upcoming two grades. Nevertheless, Saudi high schools have recently implemented a new system with more precise academic pathways to prepare students for future university degrees. These paths include computers and engineering, health, business, religion, and general science. Therefore, the question of whether a student's future academic path can be predicted arises. Moreover, it is crucial to know the most related features that affect students' academic paths, as these factors affect their lives in their early stages. Knowing these factors for students and parents and being aware of the role of these features in determining their academic and future careers is essential to prevent these drawbacks and enhance the potential of students.

What is the most appropriate ML approach to solve this problem? ML has various applications, but the predictive DM approach is the most effective. Each instance in the dataset used by the ML techniques was displayed using the exact attributes. These features can be categorical, binary, or discrete, respectively. Nevertheless, the dataset is termed unsupervised if it has no labels or correct outputs, unlike supervised ML, where the dataset is labeled [38]. The objective of supervised ML was to construct a predictive model. The created classifier used a portion of the dataset, in which the label was observed, to train the model. Class labels were then assigned to the testing rows, where the value of the class label was unknown.

Thus, the classification-prediction approach was suitable for use in this study. The proposed methodology is illustrated in Fig. 2.

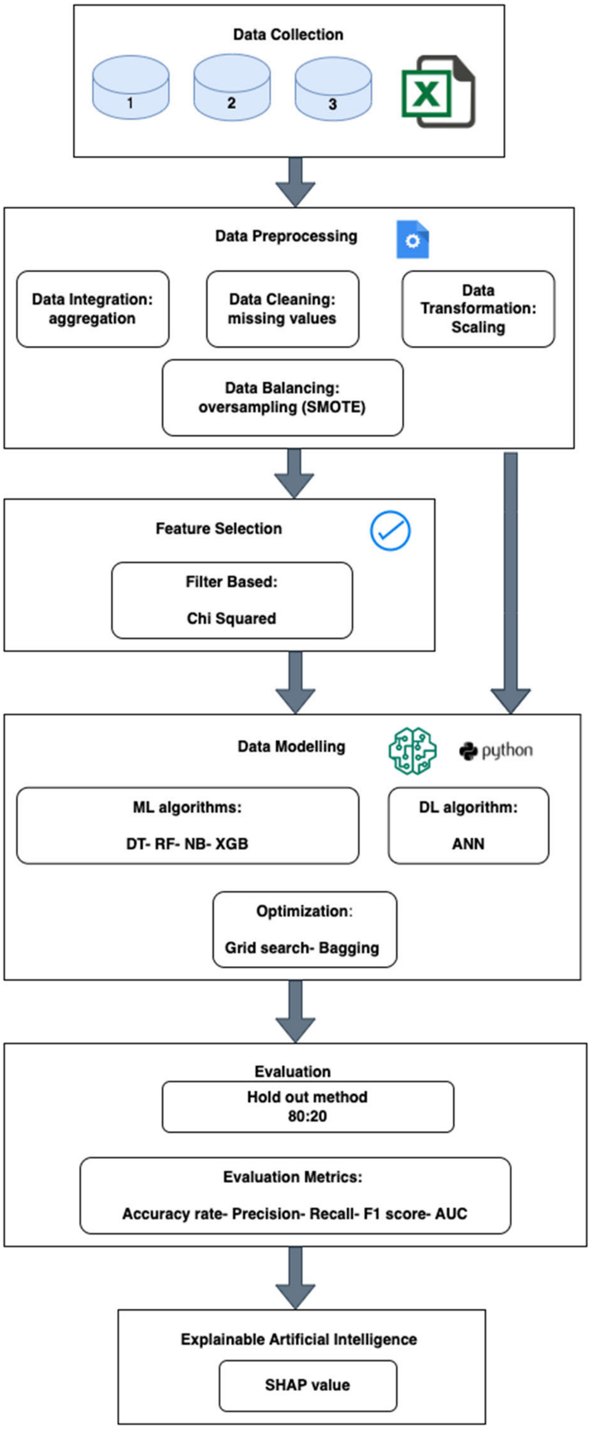

Data Collection
1
2
3
Data Preprocessing
Data Integration: aggregation
Data Cleaning: missing values
Data Transformation: Scaling
Data Balancing: oversampling (SMOTE)
Feature Selection
Filter Based: Chi Squared
Data Modelling
python
ML algorithms: DT- RF- NB- XGB
DL algorithm: ANN
Optimization: Grid search- Bagging
Evaluation
Hold out method 80:20
Evaluation Metrics: Accuracy rate- Precision- Recall- F1 score- AUC
Explainable Artificial Intelligence
SHAP value

FIGURE 2. Proposed methodology for this study.

B. DATA COLLECTION

First, we contacted the Ministry of Education in SA to request the dataset. The request was initially declined because the requested data were considered personal information and needed to be protected according to the directions of the Saudi Data and AI Authority [39]. After discussions and justifications, they accepted the request under the condition of providing coded student user IDs.

The Statistics Department and Digital Transformation Unit collaborated to produce the required dataset, which was provided in three parts in an Excel file format. The first part contained 35,406 instances and 11 features. This dataset contained the demographic features of students in the academic year 2022, who were in the 2nd grade of high school (grade 11) in the eastern province high schools of SA. In addition, it had a class label that demonstrated the chosen academic pathway for each learner.

The second part included the historical academic grades of the same students in their first year of high school (grade 10) in the 2021 academic year. The third part included historical academic grades, attendance, and behavior for the same sample in the 2020 academic year when they were in the ninth grade.

C. DATA PREPROCESSING AND DESCRIPTION

Therefore, the collected data were unsuitable for direct induction. The predominant cases included noise and missing values. Frequently encountered problems in data preprocessing include the existence of impossible, unlikely, or irrelevant input values [38].

Consequently, the collected data must be cleaned and prepared into a suitable format for processing. Hence, suitable preprocessing techniques were applied to the dataset to clean the noise, impute missing values, and address other problems. Therefore, after receiving the required datasets, they were cleaned to eliminate irrelevant features and were concluded to have 32 relevant features.

1) PREPROCESSING STEPS

Part 1: Initially, the dataset contained 11 features and 35,406 records, which included the demographic features of students in the academic year 2022 in the second grade of high school (Grade 11) in eastern provincial high schools in SA. The preprocessing steps applied in this section were as follows.

1. Drop irrelevant or unnecessary columns.
2. Convert the Arabic language into English.
3. Save the new file: EDMpart1.

There were no missing values or outliers. The features used were code, gender, location, school type, school code, and academic pathways (class label).

Part 2: The dataset contained historical academic grades for the same students in their first year of high school (Grade 10) in the 2021 academic year. It has 10 features and 419,862 records. This massive number of records can be explained by the fact that each student had 12 records for 12 courses. Hence, the preprocessing steps applied in this section were as follows.

1. Convert the Arabic language into English.

2. Drop all irrelevant or unnecessary columns.

3. Aggregate the course column with the term column in one feature.

4. Filter the entire dataset by course name.

5. Build a new table for the code, course name, and results in a new sheet.

6. Drop the course name column.

7. Change the result column name into the filtered course name.

8. Build 12 new tables for each course.

9. Merge all 12 new tables using the power query editor in Excel and use the code column as the primary key to link the 12 tables.

10. Save file name: EDMpart2.

11. Merge EDMpart1 and EDMpart2 using the power query editor in Excel.

12. Save new file name; EDMpart2 combined.

Part 3 had historical academic grades for the same sample in the 2020 academic year when they were in the ninth grade, in addition to attendance and behavior. It had nine features and 510,294 records. This massive number of records can be explained by the fact that each student has approximately 16 records for the 16 courses. Hence, the preprocessing steps applied in this section are as follows.

1. Convert the Arabic language into English.

2. Drop all irrelevant or unnecessary columns.

3. Filter the whole dataset by course name.

4. Build a new table for the code, course name, and results in a new sheet.

5. Drop the course name column.

6. Change the result column name into the filtered course name.

7. Build 13 new tables for each course.

8. The 13 new tables were merged using the Power Query Editor in Excel. Notably, the code column was used as the primary key to link the 13 tables.

9. Save file name: EDMpart3.

10. Merge EDMpart2 combined and EDMpart3 using the power query editor in Excel.

11. Save the final file name: EDMfull.

12. Some courses such as the Qur'an, family studies, and sports were not used because they were given only to males or females or were not given to all school types.

The preprocessing steps are presented in Fig. 3.

2) DATASET DESCRIPTION

Here, we provide comprehensive and descriptive tables describing all features used in this study. In addition, we presented a solution for handling imbalanced datasets. This illustrates the feature selection method used in the experiments.

a: FINAL DATASET FEATURES

Table 3 lists the final dataset features used in the MPM study.

There were missing values in some instances in the full EDM dataset. Therefore, the researchers excluded the entire row with null values rather than imputing them with the median, mode, or mean, which can introduce bias into the results. Hence, after deleting the rows with missing values, the final dataset contained 32 features and 33,878 records.

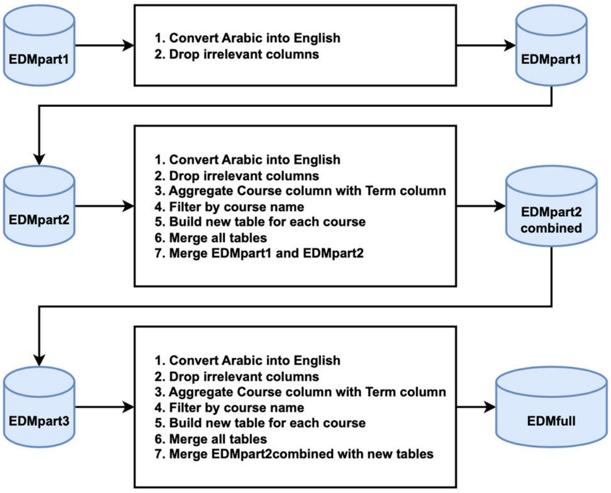

EDMpart1
1. Convert Arabic into English
2. Drop irrelevant columns
EDMpart1
EDMpart2
1. Convert Arabic into English
2. Drop irrelevant columns
3. Aggregate Course column with Term column
4. Filter by course name
5. Build new table for each course
6. Merge all tables
7. Merge EDMpart1 and EDMpart2
EDMpart2 combined
EDMpart3
1. Convert Arabic into English
2. Drop irrelevant columns
3. Aggregate Course column with Term column
4. Filter by course name
5. Build new table for each course
6. Merge all tables
7. Merge EDMpart2combined with new tables
EDMfull

FIGURE 3. Preprocessing steps.

b: DESCRIPTIVE TABLES OF FEATURES

Table 4 described the demographic features. and V described the academic features while Table 6 described the geographical features used in this study. Statistical Package for the Social Sciences (SPSS) was used to calculate the P value, mean, and standard deviation.

c: IMBALANCED DATASET

The collected dataset is imbalanced and has an unequal distribution in the label class. SMOTE [40], [41] was used to address an imbalanced dataset. SMOTE is an oversampling technique that produces synthetic minority-class instances. The SMOTE algorithm selects a random sample from the minority class. It then finds the k-nearest neighbors for the chosen sample and produces a synthetic sample between the chosen instance and its KNN from the minority class. Fig. 4 illustrates the working mechanism of the SMOTE algorithm.

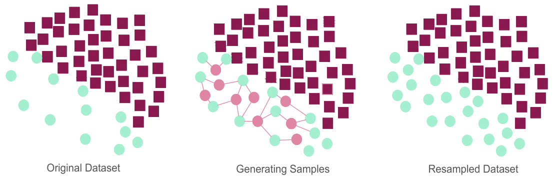


FIGURE 4. SMOTE oversampling technique.

d: FEATURE SELECTION

Feature selection uses several methods to exclude redundant attributes from a dataset, in which a minimal subset of pertinent features is employed to construct a predictive model.

TABLE 3. Features used in this study.

No.	Feature name	Description	Value	Group	Data type	Reference
1	Code (user ID)	Coded user ID	0–35406	Demographics	Numeric	[2, 8, 9, 25, 28, 29]
2	Gender	Sex	Male Female	Demographics	Nominal	[2, 9, 10, 12-15, 17, 23, 28-30]
3	Location	Geographic location for high school	12 areas	Geographics	Nominal	[10, 14, 15, 17, 22, 28, 29]
4	Type	School education type	Public Private Royal Commission of Jubail (RCJ)	Demographics	Nominal	[15, 30]
5	Schoolcode10	High School name	Statistical code	Demographics	Numeric	-
6	Schoolcode9	Secondary school name	Statistical code	Demographics	Numeric	-
7	Bio	Biology final marks for 10th grade	0–100	Academic grades	Numeric	[2, 9, 10, 13, 15-17, 20, 21, 25, 27, 29, 30, 34]
8	Chem	Chemistry final marks 10th grade	0–100	Academic grades	Numeric	
9	Comp1	Computer first term final marks 10th grade	0–100	Academic grades	Numeric	
10	Comp2	Computer second term final marks 10th grade	0–100	Academic grades	Numeric	
11	Comp3	Computer third term final marks 10th grade	0–100	Academic grades	Numeric	
12	Eng1	English language first term final marks 10th grade	0–100	Academic grades	Numeric	
13	Eng2	English language second term final marks 10th grade	0–100	Academic grades	Numeric	
14	Eng3	English language third term final marks 10th grade	0–100	Academic grades	Numeric	
15	Math1	Mathematics first term final marks 10th grade	0–100	Academic grades	Numeric	
16	Math2	Mathematics second term final marks 10th grade	0–100	Academic grades	Numeric	
17	Math3	Mathematics third term final marks 10th grade	0–100	Academic grades	Numeric	
18	Phys	Physics final marks 10th grade	0–100	Academic grades	Numeric	
19	Arabic	Arabic language final marks 10th grade	0–100	Academic grades	Numeric	
20	Rel1	Hadith final marks for 9th grade	0–100	Academic grades	Numeric	
21	Rel2	Fiqh final marks for 9th grade	0–100	Academic grades	Numeric	
22	Rel3	Tawheed final marks for 9th grade	0–100	Academic grades	Numeric	
23	Rel4	Tafseer final marks for 9th grade	0–100	Academic grades	Numeric	
24	Scie	Sciences final marks for 9th grade	0–100	Academic grades	Numeric	
25	Soci	Social studies final marks for 9th grade	0–100	Academic grades	Numeric	
26	Math9	Mathematics final marks for 9th grade	0–100	Academic grades	Numeric	
27	Comp9	Computer final marks for 9th grade	0–100	Academic grades	Numeric	
28	Eng9	English language final marks for 9th grade	0–100	Academic grades	Numeric	
29	Art9	Art final marks for 9th grade	0–100	Academic grades	Numeric	
30	Behv	Behavior final marks for 9th grade	0–100	Academic grades	Numeric	[2, 10, 16, 21]
31	Att	Attendance final marks for 9th grade	0–100	Academic grades	Numeric	
32	Class	Academic pathways in 11th grade	Five academic pathways	Academic information	Nominal	-

TABLE 4. The demographic features description.

No.	Feature name	Feature type	General (0)	Health (1)	Religion (2)	Engineering (3)	Business (4)	Overall	P-value
			n	n	n	n	n	n	
1	Gender	Demographic							0.024
	Male (0)		15,803	249	43	400	92	16,587	
	Female (1)		16,463	181	46	169	432	17,291	
								33,878	
2	Type	Demographic							-0.02
	Private (1)		3,259	38	0	131	1	3,429	
	Public (2)		27,693	392	89	438	523	29,135	
	RCJ (3)		1,314	0	0	0	0	1,314	
								33,878	

In this study, a filter-based method using the chi-square test was used to include the top ten most significant features and exclude all other features. The experimental results demonstrated that the use of the top ten features decreased the accuracy rate of the predictive models. Therefore, we used all the available features in the dataset.

D. DATA MODELING

This study was conducted to implement supervised ML classifiers and a DL technique to build models for predicting academic pathways of high school students in Saudi Arabia. DT, NB, XGB, ANN, and RF were the DM techniques used in this study. Here, we briefly describe the ML and DL algorithms used in the MPM.

As mentioned earlier, various ML algorithms are used to model data in the literature. The literature review reveals that RF, the bagging classifier (BC), XGB, DT, and NB achieve the highest accuracy rates. The RF algorithm achieved superior accuracy (99.5%) in the Bujang et al. study [25] and 97% in the Sokkhey et al. study [11]. The BC and XGB algorithms delivered their highest functionality in the Sahlaoui et al. study [23], achieving a 99.4% accuracy rate. Furthermore, the DT algorithm in the Fernandez et al. study [18] achieved a 96.5% accuracy rate, while the NB classifier in the study by Abu Amra and Maghari [14] achieved 93.17%.

1) DECISION TREE

The decision Tree (DT) [42] is a supervised algorithm. The DT learning technique is used in statistics, DM, and ML. In this formalism, classification or regression DT is used as a prediction model to draw conclusions regarding a set of observations. DTs are among the most prevalent ML methods, owing to their understandability and simplicity.

The primary algorithm utilized for constructing decision trees is known as ID3, developed by J. R. Quinlan [42]. This algorithm adopts a top-down, greedy approach, systematically exploring the available branches without any backtracking. The ID3 algorithm utilizes the concepts of entropy and information gain in order to generate a decision tree.

Entropy formula for one attribute is represented in Equation (1) and Entropy formula for two attribute is represented in Equation (2). In addition, Information Gain formula illustrated in Equation (3)

$$E\left(S\right)=\sum\nolimits_{i=1}^{c}-p_{i}log_{2}p_{i} \tag{1}$$

$$E\left(S\right)=\sum\nolimits_{c\epsilon X}-P\left(c\right)E(c) \tag{2}$$

Information Gain formula represented in Equation (3):

$$Gain\left(T,X\right)=Entropy\left(T\right)-Entrophy(T,X) \tag{3}$$

2) XTREME GRADIENT BOOSTING

Chen [43] introduced Xtreme Gradient Boosting (XGB) as a research project for a distributed (deep) ML community (DMLC) group. Currently, XGB offers package implementations for Java, Scala, Julia, Perl, and other languages as well as Python and R packages. Although the XGBoost model frequently achieves greater accuracy than a single DT, it does so at the expense of the inherent interpretability of DTs.

The mathematical formulas represented in Equations (4) and (5).

In the context of a given dataset (A, B):

$$F_{k}\left(A\right)=F_{k-1}\left(A\right)+\alpha_{k}h_{k}(A,r_{k-1}) \tag{4}$$

$$\arg min_{\alpha}=\sum\nolimits_{i=1}^{k}L(\mathrm{B}_{,},F_{i-1}\left(A_{i}\right)+\alpha h_{i}(A_{i},r_{i-1})) \tag{5}$$

3) RANDOM FOREST

The Random Forest (RF) was introduced by Breiman [44] in 2001. In this algorithm, several DTs were built during the training phase of the RF learning approach, which was used for classification, regression, and other tasks. Random decision forests amend the tendency of DTs to overfit their training set. Although they frequently outperformed DTs, gradient-boosted trees were more accurate than the RF trees.

TABLE 5. The academic feature description.

No.	Feature name	Feature type	Mean	Std. Deviation	P-value
3	BIO	Academic grades	89.29	10.804	0.042
4	CHEM	Academic grades	89	11.825	0.031
5	COMP11	Academic grades	93.74	7.981	0.024
6	COMP12	Academic grades	94.64	7.522	0.025
7	COMP13	Academic grades	93.96	8.281	0.029
8	ENG11	Academic grades	86.24	12.307	0.029
9	ENG12	Academic grades	87.26	12.845	0.036
10	ENG13	Academic grades	88.68	12.177	0.021
11	MATH11	Academic grades	84.92	12.707	0.028
12	MATH12	Academic grades	84.40	13.672	0.024
13	MATH13	Academic grades	86.16	13.603	0.036
14	PHYS	Academic grades	87.90	11.188	0.024
15	ARABIC	Academic grades	90.82	10.203	0.036
16	REL1	Academic grades	94.87	7.864	0.029
17	REL2	Academic grades	94.38	7.919	0.016
18	REL3	Academic grades	94.70	7.841	0.019
19	REL4	Academic grades	48.93	9.954	0.025
20	ATT	Academic grades	99.80	1.337	0.011
21	BEHV	Academic grades	100	0.101	0.003
22	SCIEN	Academic grades	90.53	11.207	0.020
23	SOCI	Academic grades	93.39	9.629	0.008
24	MATH9	Academic grades	87.14	13.126	0.013
25	COMP9	Academic grades	93.41	9.217	0.027
26	ENG9	Academic grades	89.01	12.109	0.032
27	ART9	Academic grades	97.27	7.314	0.002

TABLE 6. The geographical features description.

No.	Feature name	Feature type	P-value
28	Schoolcode9	Numeric	-0.016
29	Schoolcode10	Numeric	-0.034
30	Location	Nominal	-0.039

Nevertheless, the data features can affect the extent to which they work.

In the feature selection stage, the RF has a significant effect and distinct behavior. This method considers the forecast for each tree and selects the prediction with the highest number of votes. Equation (6) illustrates the RF function.

$$RFfA_a = \frac{\Sigma y\ \in all\ trees\ norm\, f\ A_{ay}}{T}, \tag{6}$$

where

RFfA sub(a) = the significance of feature x computed from all trees in the RF model

normfA sub(ay) = the importance of the standardized feature of a in tree y

T = the overall number of trees.

In the MPM study, RF was used to list the top ten essential features.

4) NB

NB [45] is a supervised ML approach that uses labeled data to train its models. This approach simplifies the computation by relying on class conditional assumptions. Therefore, it is a simple, supervised ML algorithm. NB is a statistical classifier approach that employs Bayes' theorem, which yields the following equation for the conditional probability of an event for a given b. Equations (7) and (8) illustrate Bayes' theorem:

$$P(a/b) = \frac{P\left(b/a\right) P\left(a\right)}{P\left(b\right)}, \tag{7}$$

where

P(a/b) = conditional probability of a given b

P(b) = probability of event b

P(a) = probability of event a

P(b/a) = conditional probability of b given a

$$P\left(a/b\right) = P\left(b1/a\right) \times P\left(b2/a\right) \times \ldots \times P\left(bm/a\right) \times P(a) \tag{8}$$

5) ANN

ANNs are computer programs driven by biology that imitate how the human brain processes data. ANNs acquire information by identifying data patterns, relationships, and learning from experience. An ANN comprises several individual units, artificial neurons, or processing elements interconnected by weights that compensate for the neural structure and are arranged in layers. The power of neural computation is acquired from the networked connections of neurons [46]. In the ANN structure, the training samples pass through the network and the output is obtained by comparing the network with the actual output. Subsequently, the error was used to change the weight of the neuron. Thus, the error gradually increases to achieve better results using the backpropagation algorithm [47]. The ANN hyperparameter values used in this experiment are presented in Table 7. The backpropagation algorithm is illustrated in Fig. 5.

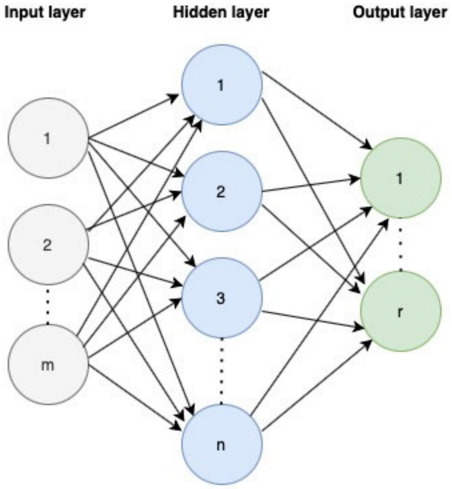


FIGURE 5. Backpropagation algorithm.

E. OPTIMIZATION

In this study, DT and RF performed well without optimization. Conversely, for the XG Boost, the hyperparameters must be tuned to obtain better results. Therefore, the grid search method was used for hyperparameter tuning, and the three hyperparameters were tuned. In addition, the bagging classifier was used to optimize the poor results for NB, and the hyperparameter values were modified, as listed in Table 8. Although the accuracy rate and other evaluation metrics increased after applying the bagging method, the overall results remained poor. Although NB performs well in the literature, it is noteworthy that it does not work well in numeric datasets, such as those used in this study; however, it performs well in categorical datasets.

TABLE 7. The ANN model's hyperparameters vlaues.

Classifier	Hyperparameters	Values
ANN	Hidden_layer_sizes	100
	Activation	Relu
	Solver	Adam
	Learning rate	0.2

TABLE 8. The XG boost and NB bagging hyperparameters values.

Classifier	Hyperparameters	Values
XG Boost	N estimators	[100,200]
	Max_depth	[8–10]
	Learning_rate	[0.001, 0.01, 0.1]
NB-Bagging	Base estimators	GussianNB
	N estimators	100

F. VERIFICATION AND EVALUATION

1) MODEL EVALUATION METHOD

In this study, all the predictive models were evaluated using the 80:20 holdout method. The dataset was divided into training and testing sets of 80% and 20%, respectively.

2) MODEL EVALUATION METRICS

Evaluation metrics are statistical measurement tools used to evaluate the classifier characteristics. This plays a vital role in obtaining an optimal classifier during the training. Thus, selecting appropriate evaluation metrics is important [48]. Therefore, this experiment evaluates several characteristics to achieve an optimal classifier. The classification problems were evaluated using a confusion matrix [48], as summarized in Table 9.

Accuracy and error rates are used to evaluate the general ability of a trained classifier; however, the accuracy metric is most used in classification problems [48]. It measures the truly predicted observations, whether positive or negative, divided by the total number of observations. The accuracy rate was calculated using (9): Conversely, the error rate complements the accuracy rate. It measures the misclassified predictions over the total number of observations. The error rate was calculated using (10):

$$AccuracyRate = \frac{TP+TN}{TP+FP+FN+TN} \tag{9}$$

$$ErrorRate = \frac{FN+TN}{TP+FP+FN+TN} \tag{10}$$

Precision is the quality of exactness and is a measure of statistical variability. Precision was calculated according to (11).

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

Recall (also known as sensitivity) is the true predicted positive divided by the true positive and false positive values.

TABLE 9. Confusion matrix for binary classification.

	Predicted positive	Predicted negative
Actual positive	True Positive (*TP*)	False Negative (*FN*)
Actual negative	False Positive (*FP*)	True Negative (*TN*)

It is calculated using (12).

$$Recall = \frac{TP}{TP + TN} \tag{12}$$

Furthermore, the F1-score is the weighted average of recall and precision, as shown in (13). This is more valuable than the accuracy in the case of an uneven class distribution.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{13}$$

The AUC is a prevalent ranking metric used to build an optimized ML model. The AUC was also used to compare the ML algorithms. Huang and Ling [49] theoretically and empirically proved that the AUC metric is better than accuracy [48]. The AUC can be calculated for binary-class problems as shown in (14).

$$AUC = \frac{S_{p-N_p(N_n+1)/2}}{N_p N_n} \tag{14}$$

where S_p is the total number of positive instances ranked and N_p and N_n are the positive and negative instances, respectively.

In summary, six evaluation metrics are recommended for use in this experiment to evaluate and optimize the predictive model, except for the error rate, because the accuracy rate was sufficient. Although previous equations were designed for binary classification, multiclass classifiers were used in this experiment. Thus, the new formulae for the multiple classifiers calculate the average values for each class.

G. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

The ML model results can be interpreted and explained to provide a more comprehensive understanding. In some cases, the interpretability of ML models is more important than their accuracy [23], [35].

The SHAP framework is an XAI application. Lundberg and Lee [50] proposed this concept. The SHAP value technique applies cooperative game principle to calculate the average marginal contribution of each feature over all potential combinations. To calculate the SHAP values, the approach examines all potential feature combinations and assesses their impact on the model's output. This enables a deep knowledge of how each attribute contributes to the prediction while accounting for both individual and interacting impacts.

The SHAP value is an XAI method that can enhance the transparency and interpretability of ML predictive models because it helps to determine the main reason for the decrease in the accuracy of a model. This powerful visualization tool can provide deep insight into the attributes that affect the model and can be immediately interpreted by nontechnical users [38]. In summary, the Python library demonstrates the contribution of each attribute to the prediction of the model. Hence, the SHAP values were calculated for the RF model in this study.

These experiments were performed on an iMac Retina 5 K with a 3.7 GHz 6-Core Intel Core i5 processor, 8 GB of memory, and a macOS Big Sur operating system. The software used in this study was "Colaboratory". It is a free Jupyter notebook environment on the cloud that allows users to write and execute Python code in their browsers and access GPUs with zero charges. Microsoft Excel was used in the preprocessing phase to integrate, aggregate, and clean the data. SPSS was used to analyze the dataset features and calculate the mean, standard deviation, and P value for each feature.

IV. RESULTS AND DISCUSSION

The major objective of this study was to create multiple ML predictive models and select the most effective one for predicting the academic pathways of students in Saudi high schools. Therefore, to answer the first question of this study five experiments were conducted.

A. RESULTS FOR FEATURE SELECTION

Experiment 1: Predictive models were designed using feature selection and filter-based methods (chi-squared test). This method was used to minimize the number of features from 31 to 10. The dataset was minimized to 5000 samples (approximately 15% of the instances). SMOTE oversampling method was used to balance the datasets. The experimental results are listed in Table 10.

In this experiment, five classifiers were used to develop predictive models. Four of them succeeded in predicting academic pathways. However, the RF outperformed the other models, achieving an accuracy rate of 98%. Additionally, it outperformed the other models in terms of the AUC, F1-score, recall, and precision. XGB initially scored 90% of the first four evaluation metrics. Therefore, we applied a grid search for hyperparameter tuning, resulting in an increase of 7% in the same metrics and a 1% increase in the AUC score.

The optimized XGB also scored a very high AUC, which was almost the same as that of the RF, but less than 1% in the accuracy rate and other evaluation metrics. Moreover, DT and ANN achieved a 96% accuracy rate, precision, recall, and F1-score; however, ANN outperformed the DT model for the AUC score.

TABLE 10. Results for experiment 1.

Model	Accuracy	Precision	Recall	F1-score	AUC
DT	96%	96%	96%	96%	97.5%
RF	98%	98%	98%	98%	99.8%
NB	40%	46%	40%	35%	72.8%
NB-bagging	44.11%	61%	44%	37%	72.88%
XGB	90.22%	90%	90%	90%	98.33%
XGB- tuning	97%	97%	97%	97%	99.74%
ANN	96.08%	95.99%	96.08	95.99%	99.53%

TABLE 11. Results for experiment 2.

Model	Accuracy	Precision	Recall	F1-score	AUC
DT	97%	97%	97%	97%	98.35%
RF	98%	98%	98%	98%	99.85%
NB	38%	52%	38%	33%	70.86%
NB-bagging	38%	52%	38%	33%	70.86%
XGB	81%	81%	81%	80%	95.1%
XGB-tuning	95%	95%	95%	95%	99.5%
ANN	96.1%	96.2%	96.1%	96%	99.6

Conversely, NB exhibited poor results in most evaluation metrics, except for AUC, which scored almost 73%. Therefore, a bagging method was used to optimize the model. Consequently, the accuracy rate, F1-score, recall, and precision scores increased; however, they remained weak.

Experiment 2: Predictive models were developed using feature selection and filter-based methods (chi-squared test). This method was used to minimize the number of features from 31 to 10. All instances in the dataset were used. In addition, the SMOTE oversampling method was used to balance the dataset. The results are presented in TABLE 11.

In this experiment, RF outperformed the other models in all evaluation metrics. Notably, using the entire dataset in this experiment did not change the scores achieved for each evaluation metric in the RF and ANN models. However, it increased the accuracy, precision, F1-score, and recall of the DT model by 1%. However, the use of the entire dataset negatively affected the NB model by decreasing the scores for all the evaluation metrics. In addition, the bagging method does not affect the NB model. The XGB model scores decreased by 9% in accuracy rate, precision, and recall and they decreased by 10% and 3% in F1-score and AUC, respectively. A grid-search method was utilized to optimize the XGB model, which increased the accuracy rate of the model by 14%.

B. RESULTS FOR ALL THE FEATURES USED

Experiment 3: In this experiment, Predictive models were developed using full features and a sample of instances with a sample size of 5000 (approximately 15% of the instances in the dataset). The SMOTE oversampling method was used in this experiment. The results are presented in Table 12.

As shown, using all features in the dataset did not affect the DT model; however, it had a positive effect on all other models. Although the NB and NB bagging models exhibited an increase in all metrics, their results remained weak and were considered to have failed. In addition, the XGB model benefited from using all features. The XGB tuning model and RF outperformed each other by scoring 99% in accuracy, precision, recall, and F1-score, and above 99.9% for AUC. In the last model, the ANN was improved by 1% and scored 99.7% for AUC and 97% for all other metrics.

Experiment 4: In this experiment, we developed predictive models using all available features in the dataset and a sample size of 5000 (approximately 15% of the instances of the dataset). The SMOTE oversampling method was not used. The experimental results are presented in Table 13. (Note: Weighted average was used, but the macro average was not).

This experiment revealed the effect of using an imbalanced dataset without an oversampling method. The macro-average of each evaluation metric is different from the weighted average. The macro-average returns the average without considering the proportion of each class on the label. Conversely, the weighted average returns the average by considering the proportion of each class on the label. Therefore, a weighted average was used in the experiments.

In contrast to Experiment 3, all models were affected by using of an imbalanced dataset. Nevertheless, the RF and XGB models outperformed all other models by achieving an

TABLE 12. Results for experiment 3.

Model	Accuracy	Precision	Recall	F1-score	AUC
DT	97%	97%	97%	97%	98.1%
RF	99%	99%	99%	99%	99.94%
NB	44%	61%	44%	44%	77.56%
NB-bagging	44%	60%	44%	37%	77.6%
XGB	93%	93%	93%	93%	99.1%
XGB-tuning	99%	99%	99%	99%	99.97%
ANN	97%	97%	97%	97%	99.72%

TABLE 13. Results of experiment 4.

Model	Accuracy	Precision	Recall	F1-score	AUC
DT	91%	92%	91%	91%	56.8%
RF	95.1%	92%	95%	93%	69.6%
NB	10%	95%	10%	14%	68.7%
NB-bagging	12%	95%	12%	18%	70.8%
XGB	95%	90%	95%	93%	88.7%
XGB-tuning	95%	92%	95%	93%	82.5%
ANN	93.6%	90.6%	93.6%	92%	61.5%

TABLE 14. Results of experiment 5.

Model	Accuracy	Precision	Recall	F1-score	AUC
DT	98%	98%	98%	98%	98.5%
RF	99%	99%	99%	99%	99.97%
NB	37%	55%	37%	28%	72.2%
NB-bagging	37%	55%	37%	28%	72.2%
XGB	82%	83%	82%	82%	96.1%
XGB-tuning	98%	98%	98%	98%	99.9%
ANN	97.5%	97.5%	97.5%	97.5%	99.8%

accuracy rate of 95%; however, XGB scored the highest AUC and outperformed the RF model.

In summary, XGB outperforms all other models in the case of an imbalanced dataset; however, it consumes extended running time compared to RF.

Experiment 5: In this experiment, we developed predictive models using all available features and instances of the dataset (100% of the instances of the dataset). The SMOTE oversampling method was also used. The experimental results are listed in Table 14.

The setting for this experiment was to use all dataset features and instances in addition to the resampling method SMOTE, which generates equal numbers of classes. We established that the RF model was superior in achieving a 99% accuracy rate, precision, recall, and F1-score and further scored 99.97% in the AUC metric. This was followed by the XGB tuning model, which scored 99.9% for AUC and 98% for all other evaluation metrics. Similarly, the DT model scored 98.5% for AUC and 98% for all other metrics. The ANN model scored 99.8% for AUC and 97.5% for all other metrics.

Therefore, the answer for the first research question obvious that the academic pathways for Saudi high school students in can be predicted effectively.

C. RESULTS OF THE MOST SIGNIFICANT FEATURES

The second question of this study was RQ2: What are the most relevant features affecting students' academic pathways?

RF achieved the highest accuracy rate among the other models in most experiments and was used to rank the most critical features in descending order.

As illustrated in Fig. 6, the most influential feature was the final physics marks provided for first-year high school students. Subsequently, students in their first year of high school received the final marks of the third mathematics course. This was followed by the second and first mathematics courses of the same year. Subsequently, three English classes were conducted. Subsequently, three computer classes were conducted. This was followed by the chemical and biological studies. The high school's name came immediately after all courses in the first year.

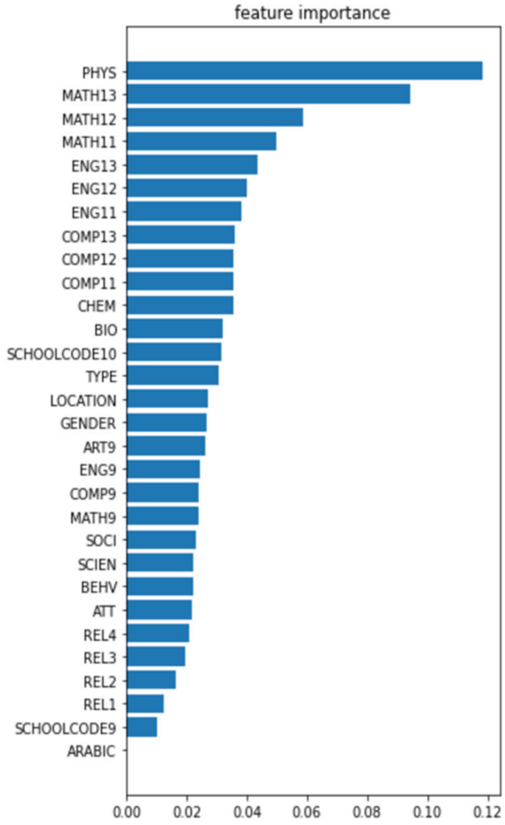


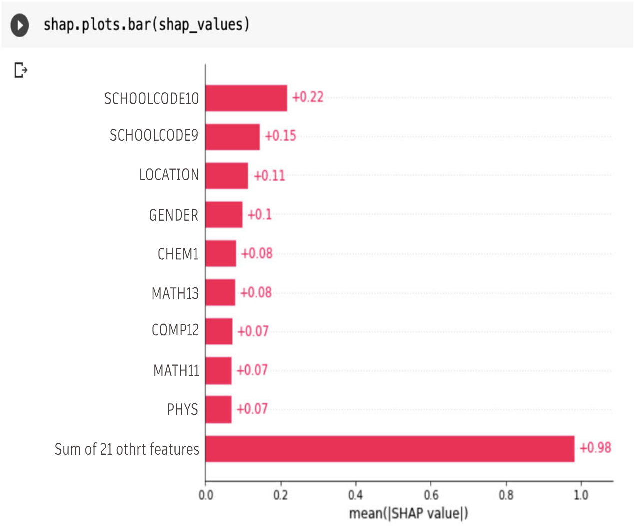
FIGURE 6. **Feature importance results for RF model.**

D. RESULTS ON THE XAI (SHAP VALUES)

ML models are often represented by black boxes. Therefore, we provide the results of utilizing SHAP values to explain and interpret the ML models and further count the contribution of each feature to the predictive process. Therefore, this approach was used to explain the RF model, which is the best predictive model among the other models.

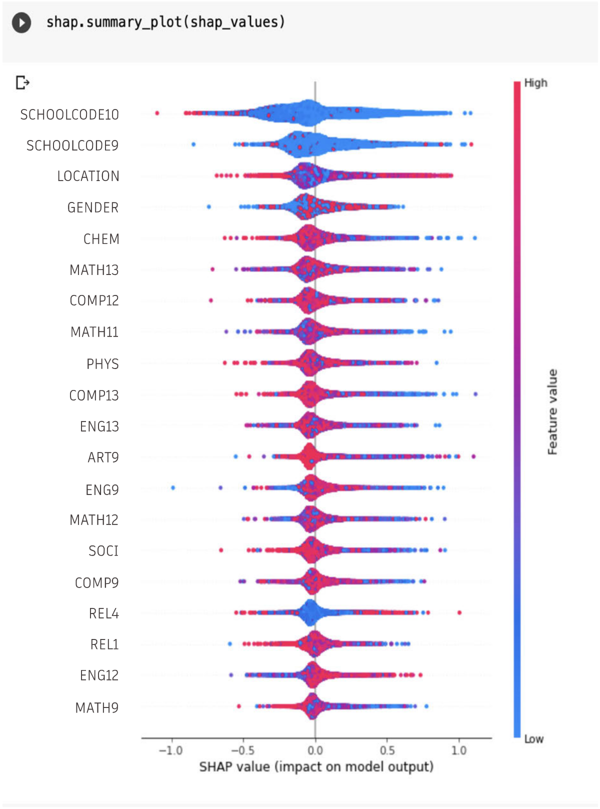
Fig. 7 shows the SHAP plot of the RF model. It illustrates the features are ordered from those with the highest to lowest influence on the prediction. It counts the absolute SHAP values for each feature contribution without determining whether the contribution is positive or negative.

Fig. 8 shows the summary plot and Fig. 9 shows the bee swarm count of the contribution of each feature with a positive or negative impact. Thus, the features are ordered according to their influence on the prediction and further illustrate how higher or lower feature values affect the result. The small dots indicate a single observation, and the horizontal axis represents the SHAP values.



```
shap.plots.bar(shap_values)
```

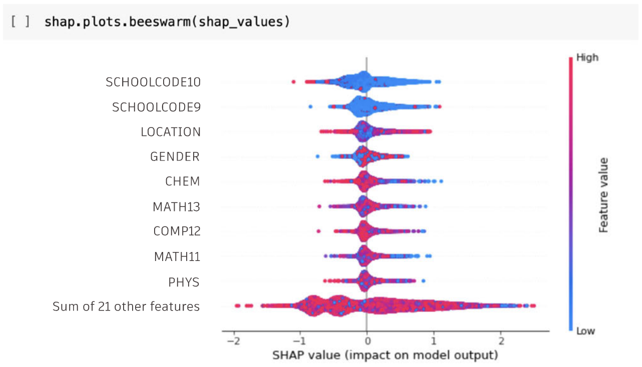
FIGURE 7. **SHAP plots bar for the RF model.**



```
shap.summary_plot(shap_values)
```

FIGURE 8. **SHAP summary plot for the RF model.**

In this study, the experimental results demonstrated that the SCHOOLCODE10 feature positively affected the prediction and was an important feature that contributed to the prediction



```
[ ]  shap.plots.beeswarm(shap_values)
```

FIGURE 9. SHAP bee swarm for the RF model.

process. SCHOOLCODE10 provides the name of the high school for each student. Hence, the name of the school in which each learner was enrolled was highly related to and influenced the academic pathways of the students.

The second most influential feature was SCHOOLCODE9, which positively affected the prediction. These results indicate that middle school names positively affect the prediction of students' academic pathways.

LOCATION had the third highest impact on the prediction, and GENDER was the fourth most influential feature. Geographic and demographic features were the dominant factors affecting the students' academic pathways. This is followed by academic features that have less impact on prediction. Therefore, MATH13 and CHEM had the same impact on the prediction process. This was followed by COMP12, PHYS, and MATH11, which had the same negative impacts on the prediction process.

The prediction model was easier to understand and more interpretable when the SHAP value technique was applied. The report provided essential insight on the relative importance of different features, with special attention to how school location influences Saudi high school students' academic paths. Educational interventions and policies that aim to improve student success in higher education and improve educational outcomes can benefit greatly from the knowledge contained in these findings.

V. CONCLUSION

In this study, we demonstrated that the academic pathways of Saudi high school students can be effectively predicted. The study involved six steps: defining the problem, collecting the required dataset, preparing and preprocessing the dataset, and modeling and optimizing to verify and deploy the model. Furthermore, this study extends the research by applying XAI to interpret the highest model to better understand and enhance the transparency of the predictive model. Hence, five DM algorithms were used to construct predictive models to answer the research questions.

The experimental results demonstrate that using the SMOTE oversampling method improves the evaluation metrics of most classifiers. Moreover, using the feature selection method reduced the evaluation metrics for most classifiers, whereas utilizing all available instances increased the evaluation metrics for most of the built classifiers. Therefore, the ideal settings for the experiments were to use the SMOTE resampling method, all available features, and all the data in the dataset.

The RF and XGB tuning models outperformed the other classifiers when using full features, SMOTE resampling method, and a sample size of 5000. RF and XGB outperformed each other when an imbalanced dataset was used, achieving an accuracy rate of 95%. RF also outperformed the ideal case by achieving a 99% accuracy rate, precision, recall, and F1-score and it achieved the highest AUC metric by scoring 99.97%.

The RF model was used to identify the top ten crucial features that affect academic pathways for high school students in Saudi Arabia. We established that physics courses were the most influential features in high schools, followed by mathematics, English, and computer courses.

Furthermore, the SHAP values were used to explain and interpret the superior predictive model. The experimental results revealed that the name of the high school in which each student enrolled contributed immensely to the prediction process. In summary, geographic and demographic features were the dominant features that contributed more to the prediction process than academic grades.

In this study, a predictive model with excellent accuracy was developed. Additionally, we interpreted and explained this to enhance transparency. Therefore, we recommend that decision makers in the Ministry of Education in Saudi Arabia utilize this model, as it can help in the planning and development of processes by predicting the future academic pathways of students. Furthermore, it could accelerate the digital transformation of educational systems as part of Saudi Vision 2030.

A limitation of this study is that the dataset used covered only eastern providence. Therefore, we recommend continuing the research using a dataset covering all high schools in SA. Additionally, we encourage parents, guardians, and students to pay considerable attention to physics, mathematics, and English courses in tenth grade, as they have significant influence on students' academic pathways in Saudi high schools.

ETHICS

The study design was approved by the ethical board of Imam Abdulrahman bin Faisal University.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Education, Saudi Arabia, and its employees for their support and providing the dataset. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. This study received no funding. There are no conflicts of interest to declare.

REFERENCES

[1] *MO Education*. Accessed: Aug. 18, 2022. [Online]. Available: https://twitter.com/moe_gov_sa?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

[2] A. Qazdar, B. Er-Raha, C. Cherkaoui, and D. Mammass, "A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco," *Educ. Inf. Technol.*, vol. 24, no. 6, pp. 3577–3589, Nov. 2019, doi: 10.1007/s10639-019-09946-8.

[3] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern., C*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[4] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[5] A. V. Manjarres, L. G. M. Sandoval, and M. J. S. Suárez, "Data mining techniques applied in educational environments: Literature review," *Digit. Educ. Rev.*, vol. 33, pp. 235–266, Jun. 2018.

[6] C. Romero and S. Ventura, "Data mining in education," *WIREs Data Mining Knowl. Discovery*, vol. 3, no. 1, pp. 12–27, Jan. 2013, doi: 10.1002/widm.1075.

[7] M. F. M. Mohsin, C. F. Hibadullah, N. M. Norwawi, and M. H. A. Wahab, "Mining the student programming performance using rough set," in *Proc. IEEE Int. Conf. Intell. Syst. Knowl. Eng.*, Nov. 2010, pp. 478–483, doi: 10.1109/ISKE.2010.5680824.

[8] J. Kuzilek, Z. Zdrahal, and V. Fuglik, "Student success prediction using student exam behaviour," *Future Gener. Comput. Syst.*, vol. 125, pp. 661–671, Dec. 2021, doi: 10.1016/j.future.2021.07.009.

[9] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 928, no. 3, 2020, Art. no. 032019, doi: 10.1088/1757-899X/928/3/032019.

[10] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, "Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation," *Soft Comput.*, vol. 24, no. 24, pp. 18477–18487, Dec. 2020, doi: 10.1007/s00500-020-05075-4.

[11] P. Sokkhey, S. Navy, L. Tong, and O. Takeo, "Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in Cambodia," *IEIE Trans. Smart Process. Comput.*, vol. 9, no. 3, pp. 217–229, Jun. 2020.

[12] F. Saleem, Z. Ullah, B. Fakieh, and F. Kateb, "Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning," *Mathematics*, vol. 9, no. 17, p. 2078, Aug. 2021, doi: 10.3390/math9172078.

[13] J. L. Harvey and S. A. Kumar, "A practical model for educators to predict student performance in K-12 education using machine learning," in *Proc. IEEE Symp. Series Comput. Intell. (SSCI)*, Dec. 2019, pp. 3004–3011, doi: 10.1109/SSCI44817.2019.9003147.

[14] I. A. Abu Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in *Proc. 8th Int. Conf. Inf. Technol. (ICIT)*, May 2017, pp. 909–913, doi: 10.1109/ICITECH.2017.8079967.

[15] S. Priya, T. Ankit, and D. Divyansh, "Student performance prediction using machine learning," in *Advances in Parallel Computing Technologies and Applications*. Amsterdam, The Netherlands: IOS Press, 2021, pp. 167–174.

[16] K. Sunday, P. Ocheja, S. Hussain, S. S. Oyelere, B. O. Samson, and F. J. Agbo, "Analyzing student performance in programming education using classification techniques," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 15, no. 2, p. 127, Jan. 2020, doi: 10.3991/ijet.v15i02.11527.

[17] M. Ahmed and A. Mahmood, "An empirical study of machine learning algorithms to predict students'grades," *Pakistan J. Sci.*, vol. 70, no. 1, pp. 91–96, 2018.

[18] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, "Application of machine learning in predicting performance for computer engineering students: A case study," *Sustainability*, vol. 11, no. 10, p. 2833, May 2019, doi: 10.3390/su11102833.

[19] A. Tarik, H. Aissa, and F. Yousef, "Artificial intelligence and machine learning to predict student performance during the COVID-19," *Proc. Comput. Sci.*, vol. 184, pp. 835–840, Jan. 2021, doi: 10.1016/j.procs.2021.03.104.

[20] P. Kenekayoro, "An exploratory study on the use of machine learning to predict student academic performance," *Int. J. Knowl.-Based Organizations*, vol. 8, no. 4, pp. 67–79, Oct. 2018, doi: 10.4018/ijkbo.2018100104.

[21] E. Wakelam, A. Jefferies, N. Davey, and Y. Sun, "The potential for student performance prediction in small cohorts with minimal available attributes," *Brit. J. Educ. Technol.*, vol. 51, no. 2, pp. 347–370, Mar. 2020, doi: 10.1111/bjet.12836.

[22] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a web-based prediction system for students' academic performance," *Data*, vol. 7, no. 2, p. 21, Jan. 2022, doi: 10.3390/data7020021.

[23] H. Sahlaoui, E. A. A. Alaoui, A. Nayyar, S. Agoujil, and M. M. Jaber, "Predicting and interpreting student performance using ensemble models and Shapley additive explanations," *IEEE Access*, vol. 9, pp. 152688–152703, 2021, doi: 10.1109/ACCESS.2021.3124270.

[24] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, Aug. 2016, doi: 10.14257/ijdta.2016.9.8.13.

[25] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. Md. Ghani, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[26] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[27] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks," *IEEE Access*, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.

[28] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[29] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, "Detecting at-risk students with early interventions using machine learning techniques," *IEEE Access*, vol. 7, pp. 149464–149478, 2019, doi: 10.1109/ACCESS.2019.2943351.

[30] A. Lin, Q. Wu, A. A. Heidari, Y. Xu, H. Chen, W. Geng, Y. Li, and C. Li, "Predicting intentions of students for master programs using a chaos-induced sine cosine-based fuzzy K-nearest neighbor classifier," *IEEE Access*, vol. 7, pp. 67235–67248, 2019, doi: 10.1109/ACCESS.2019.2918026.

[31] D. Liu, Y. Zhang, J. Zhang, Q. Li, C. Zhang, and Y. Yin, "Multiple features fusion attention mechanism enhanced deep knowledge tracing for student performance prediction," *IEEE Access*, vol. 8, pp. 194894–194903, 2020, doi: 10.1109/ACCESS.2020.3033200.

[32] A. Polyzou and G. Karypis, "Feature extraction for next-term prediction of poor student performance," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 237–248, Apr. 2019, doi: 10.1109/TLT.2019.2913358.

[33] O. Iatrellis, I. K. Savvas, A. Kameas, and P. Fitsilis, "Integrated learning pathways in higher education: A framework enhanced with machine learning and semantics," *Educ. Inf. Technol.*, vol. 25, no. 4, pp. 3109–3129, Jul. 2020, doi: 10.1007/s10639-020-10105-7.

[34] O. Iatrellis, I. K. Savvas, P. Fitsilis, and V. C. Gerogiannis, "A two-phase machine learning approach for predicting student outcomes," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 69–88, Jan. 2021, doi: 10.1007/s10639-020-10260-x.

[35] D. Datamanv, "Explain your model with the SHAP values," *Towards Data Sci.*, Sep. 2019. [Online]. Available: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

[36] R. Alamri and B. Alharbi, "Explainable student performance prediction models: A systematic review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021, doi: 10.1109/ACCESS.2021.3061368.

[37] P. Arjunan, K. Poolla, and C. Miller, "EnergyStar++: Towards more accurate and explanatory building energy benchmarking," *Appl. Energy*, vol. 276, Oct. 2020, Art. no. 115413, doi: 10.1016/j.apenergy.2020.115413.

[38] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

[39] SDAIA. *Saudi Data and AI Authority*. Accessed: Nov. 20, 2022. [Online]. Available: https://sdaia.gov.sa/ar/SDAIA/about/Documents/Policies001.pdf

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[41] L. Abdi and S. Hashemi, ''To combat multi-class imbalanced problems by means of over-sampling techniques,'' *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 238–251, Jan. 2016, doi: 10.1109/TKDE.2015.2458858.

[42] J. R. Quinlan, ''Probabilistic decision trees,'' in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 1990, pp. 140–152.

[43] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, and H. Cho, ''Xgboost: Extreme gradient boosting,'' *R Package Version*, vol. 1, no. 4, pp. 1–4, Aug. 2015.

[44] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[45] D. J. Hand and K. Yu, ''Idiot's Bayes—Not so stupid after all?'' *Int. Stat. Rev.*, vol. 69, no. 3, pp. 385–398, Dec. 2001, doi: 10.1111/j.1751-5823.2001.tb00465.x.

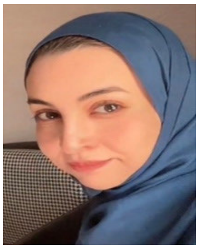
[46] S. Agatonovic-Kustrin and R. Beresford, ''Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research,'' *J. Pharmaceutical Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, Jun. 2000, doi: 10.1016/s0731-7085(99)00272-1.

[47] J. Zupan, ''Introduction to artificial neural network (ANN) methods: What they are and how to use them,'' *Acta Chim. Slovenica*, vol. 41, p. 327, Jan. 1994.

[48] M. Hossin and M. N. Sulaiman, ''A review on evaluation metrics for data classification evaluations,'' *Int. J. Data Mining Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[49] J. Huang and C. X. Ling, ''Using AUC and accuracy in evaluating learning algorithms,'' *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.

[50] S. M. Lundberg and S.-I. Lee, ''A unified approach to interpreting model predictions,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.



MAI ABDALKAREEM received the B.S. degree in computer science from Dammam University, Dammam, Saudi Arabia, in 2004. She is currently pursuing the master's degree in advanced computer science with Imam Abdulrahman bin Faisal University, Dammam. She was with the Ministry of Education, Saudi Arabia, for 12 years as a Computer Science Teacher.

NASRO MIN-ALLAH received the Ph.D. degree in CS from the Graduate School of the Chinese Academy of Sciences, in 2007. He completed his postdoctoral research from MIT, in 2014. He is currently a Professor in computer science with the College of Computer Science and Information Technology, University of Dammam, Saudi Arabia. He was with the SuperTech Group, MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), from September 2012 to June 2014, as a Visiting Scientist and taught with the Electrical Engineering and Computer Science Department, from January 2013 to May 2014. He distinguished between careers in education, research, and administration. He was an Associate Professor and the Head of the Department of Computer Science, COMSATS Institute of Information Technology (CIIT), Pakistan, from 2002 to 2012, and the Director of the Green Computing and Communication Laboratory. He was a recipient of three prestigious awards: The CIIT Golden Medallion for Innovation (CIMI-2009), Best Mobile Innovation in Pakistan (BMIP-2010), and the Best University Teacher Award in Pakistan (BUTA-2011).